(19) JAPANESE PATENT OFFICE

PATENT ABSTRACTS OF JAPAN

(11) Publication number: **2000174817 A**

(43) Date of publication of application: **23.06.00**

(51) Int. Cl **H04L 12/56**
H04Q 3/00
H04Q 11/04

(21) Application number: **11172584**

(22) Date of filing: **18.06.99**

(30) Priority: **08.12.98 US 98 206975**

(71) Applicant: **NEC CORP**

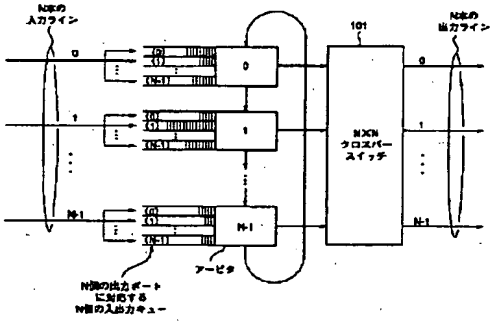
(72) Inventor: **RAMAMURTHY GOPALAKRISHNAN**
FAN RUIXUE
SMILJANIC ALEKSANDRA

(54) **METHOD AND SYSTEM FOR SCHEDULING IN EXCHANGE SYSTEM**

(57) Abstract:

PROBLEM TO BE SOLVED: To allow a title system to satisfy a tight timing requirement by selecting one output from among a set of outputs that can be available for a future time slot, corresponding to a selected input and storing the selected input and the selected output relating to the selected input in pairs as a schedule.

SOLUTION: An ultrahigh speed exchange system consists of an N_xN crossbar switch 101, and packets received from N-sets of input lines consist of cells with a fixed length entirely. Each of the N-sets of input ports has N-sets of arbiters respectively and each input port has N-sets of theoretical queues corresponding respectively to N-sets of output ports. Each of the N-sets of the arbiters (0, 1,..., N-1) receives a set of outputs of the available output ports from a preceding arbiter sequentially each time slot and selects one output port from among the available output ports according to the round robin system. The output of the selected output port is excluded from the set of outputs.



(19)日本国特許庁（ＪＰ）　(12)公開特許公報（Ａ）　(11)特許出願公開番号
特開2000－174817
（P2000－174817A）
(43)公開日　平成12年６月23日(2000.6.23)

(51)Int.Cl.'	識別記号	ＦＩ		テーマコード'（参考）
Ｈ０４Ｌ　12/56		Ｈ０４Ｌ	11/20	１０２Ｚ
Ｈ０４Ｑ　3/00		Ｈ０４Ｑ	3/00	
11/04			11/04	Ｒ

審査請求　有　　請求項の数18　ＯＬ　（全 16 頁）

(21)出願番号　特願平11－172584

(22)出願日　平成11年６月18日(1999.6.18)

(31)優先権主張番号　０９／２０６９７５

(32)優先日　平成10年12月８日(1998.12.8)

(33)優先権主張国　米国（ＵＳ）

(71)出願人　000004237
日本電気株式会社
東京都港区芝五丁目７番１号

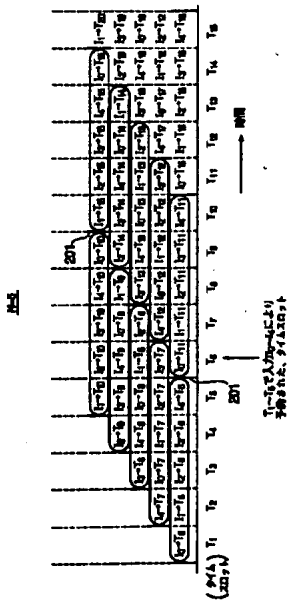
(72)発明者　ゴパラクリシナン　ラママーシー
アメリカ合衆国，ニュージャージー，
08540　プリンストン，インディペンデンス　ウエイ　４，エヌ・イー・シー・ユー・エス・エー・インク内

(74)代理人　100097157
弁理士　桂木　雄二

最終頁に続く

(54)【発明の名称】　交換システムにおけるスケジューリング方法及び装置

(57)【要約】
【課題】　超高速交換システムにおいて入出力間スケジュールを高速に確立することができる新規なスケジューリング方法及び装置を提供する。
【解決手段】　Ｎ×Ｎ交換システムにおいて、Ｎ入力の各々にＮ出力にそれぞれ対応した論理的キューと、Ｎ入力にそれぞれ対応し隣接アービタへ制御情報を送出するＮ個のアービタとが設けられている。複数のスケジューリング過程がタイミングをずらせて開始され、Ｎタイムスロットだけ未来のタイムスロットがそれぞれ決定される。開始アービタからラウンドロビン方式で順次１つのアービタが選択され、選択されたアービタは未来のタイムスロットにおける利用可能な出力の集合の中から１つの出力を選択する。選択された出力を出力集合から除外し、択された入力とそれに関連する選択された出力との組をスケジュールとしてメモリに記憶する。複数のスケジューリング過程をパイプライン処理により同時に並列処理できるために、スケジュールを高速に確立することができる。



1

【特許請求の範囲】

【請求項１】　Ｎ入力及びＮ出力（Ｎは２以上の整数）を有し、各入力は前記Ｎ出力にそれぞれ対応したＮ個の論理的待ち行列（論理的キュー）からなる交換システムにおけるスケジューリング方法において、

ａ）任意のスケジューリング過程を開始する入力に対して、当該開始入力から予め定められた数のタイムスロットだけ未来のタイムスロットを決定し、

ｂ）前記開始入力からラウンドロビン方式で順次１つの入力を選択し、

ｃ）選択された入力に転送すべきパケットが存在すれば、当該選択された入力に対して、前記未来のタイムスロットにおける利用可能な出力の集合の中から１つの出力を選択し、

ｄ）選択された出力を前記出力集合から除外し、前記選択された入力とそれに関連する前記選択された出力との組をスケジュールとして記憶する、

ステップからなることを特徴とするスケジューリング方法。

【請求項２】　複数のスケジューリング過程がパイプライン処理により同時に進行することを特徴とする請求項１記載のスケジューリング方法。

【請求項３】　Ｎが奇数の場合、前記未来のタイムスロットは前記開始入力からＮタイムスロットだけ未来に位置し、前記選択された出力を前記出力集合から除外した出力集合を次に選択されるべき入力へ転送する、ことを特徴とする請求項１記載のスケジューリング方法。

【請求項４】　Ｎ個のスケジューリング過程が１タイムスロットずつ位相をずらしながら同時に進行することを特徴とする請求項３記載のスケジューリング方法。

【請求項５】　Ｎが偶数の場合、前記未来のタイムスロットは前記開始入力から（Ｎ＋１）タイムスロットだけ未来に位置し、前記選択された出力を前記出力集合から除外した出力集合を次に選択されるべき入力へ転送する動作を（Ｎ＋１）タイムスロットのうち１回だけ１タイムスロット分遅延させる、ことを特徴とする請求項１記載のスケジューリング方法。

【請求項６】　（Ｎ＋１）個のスケジューリング過程が１タイムスロットずつ位相をずらしながら同時に進行することを特徴とする請求項５記載のスケジューリング方法。

【請求項７】　前記交換システムは、更に、マルチキャスト・パケットを格納するための別個のキューを有し、各キューは当該キューのHOLパケットのあて先を示すマルチキャストビットマップ(BM)を有し、

マルチキャスト・キューはユニキャスト・キューよりも優先されて処理され、任意の入力は前記出力集合とマルチキャストビットマップとの積集合を満たす全ての出力を選択し、

HOLマルチキャストパケットは、前記未来のタイムスロットにおいて選択された全ての出力へ送信される、

ことを特徴とする請求項１ないし６のいずれかに記載のスケジューリング方法。

2

【請求項８】　Ｎ入力及びＮ出力（Ｎは２以上の整数）を有し、各入力は前記Ｎ出力にそれぞれ対応したＮ個の論理的待ち行列（論理的キュー）からなる交換システムにおいて、

前記Ｎ入力の各々は、

到着パケットを格納して前記論理的キューを管理する入力格納手段と、

前記到着パケットの出力要求に応じて、利用可能な出力の集合の中から１つの出力を選択して前記入力格納手段へ送出するアービタ手段と、

前記アービタ手段により選択された出力とその入力との組を接続情報として格納する接続格納手段と、

前記アービタ手段を隣接するアービタ手段と協調して制御するパイプライン制御手段と、

からなり、

前記パイプライン制御手段は、

ａ）任意のスケジューリング過程を開始する入力のアービタ手段に対して、当該開始アービタ手段から予め定められた数のタイムスロットだけ未来のタイムスロットを決定し、

ｂ）前記開始アービタ手段からラウンドロビン方式で順次１つのアービタ手段を選択し、

ｃ）選択されたアービタ手段に対応する入力格納手段から出力要求があれば、当該選択されたアービタ手段により、前記未来のタイムスロットにおける利用可能な出力の集合の中から１つの出力を選択させ、

ｄ）選択された出力を前記出力集合から除外し、前記選択されたアービタ手段とそれに関連する前記選択された出力との組をスケジュールとして前記接続格納手段へ格納する、

ことを特徴とするスケジューリング装置。

【請求項９】　Ｎ入力及びＮ出力（Ｎは２以上の整数）を有するＮ×Ｎクロスバスイッチを有する交換システムにおいて、

前記Ｎ入力の各々は、

到着パケットを格納して前記Ｎ出力にそれぞれ対応したＮ個の論理的待ち行列（論理的キュー）管理する入力モジュールと、

前記到着パケットの出力要求に応じて、利用可能な出力の集合の中から１つの出力を選択して前記入力格納手段へ送出するアービタと、

前記アービタにより選択された出力とその入力との組を格納する接続メモリと、

前記アービタを隣接するアービタと協調して制御するパイプライン制御部と、

からなり、

前記Ｎ出力の各々は出力モジュールからなり、

前記パイプライン制御手段は、
ａ）複数のスケジューリング過程を順次開始するアービタに対して、当該開始アービタから予め定められた数のタイムスロットだけ未来のタイムスロットをそれぞれ決定し、
ｂ）前記開始アービタからラウンドロビン方式で順次１つのアービタを選択し、
ｃ）選択されたアービタに対応する入力モジュールから出力要求があれば、当該選択されたアービタにより、前記未来のタイムスロットにおける利用可能な出力の集合の中から１つの出力を選択させ、
ｄ）選択された出力を前記出力集合から除外し、前記選択されたアービタとそれに関連する前記選択された出力との組をスケジュールとして前記接続メモリへ格納し、前記スケジュールに従って入力モジュールのパケットを前記クロスバースイッチを通して選択された出力へ送出することを特徴とする交換システム。

【請求項１０】　Ｎ入力及びＮ出力（Ｎは２以上の整数）を有し、各入力は前記Ｎ出力にそれぞれ対応したＮ個の論理的待ち行列（論理的キュー）からなるラウンド・ロビン・グリーティ・スケジューリングプロトコルのためのクロスバースイッチにおけるタイムスロット決定方法であって、前記プロトコルの入力は全ての入出力キューの状態であり、前記プロトコルの出力はスケジュールであり、
ａ）ｉ=(定数－ｋ－1)modNに対応する入力を選択し、
ｂ）もし入力が無ければ停止し、それ以外であればラウンドロビンのやり方でi = (i + 1) mod Nにより決定される次の入力を選択し、
ｃ）集合C={（i，j）|出力jに対応する入力ｉにおいて少なくとも１個のパケットが存在する}の要素である組（ｉ，ｊ）が存在するならば、出力jを選択し、
ｄ）ステップｃ）において前記組（ｉ，ｊ）が存在しなければ、入力集合からｉを除去してステップｂ）に戻り、
ｅ）入力集合からiを、出力集合からjをそれぞれ除去し、
ｆ）前記組（ｉ，ｊ）を前記スケジュールに加えてステップｂ）に戻る、
ステップからなることを特徴とするタイムスロットの決定方法。

【請求項１１】　各タイムスロットにおいて、Ｎ個の異なるスケジュールが未来のＮタイムスロット間で同時進行するスケジューリング方法において、
ａ）ラウンドロビン方式で、スケジューリングのための入力に対して特定の未来のタイムスロットを利用可能にし、
ｂ）入力ｉによって、未来のｋ番目のタイムスロットに対する出力を選択し、
ｃ）未来のｋ番目のライムスロットに対するスケジュールを開始し、
ｄ）次の入力（ｉ＋１）ｍｏｄＮを決定し、前記ｋ番目のタイムスロットの間にパケットを自由に受信できる残りの出力を前記次の入力へ送出する、
ステップからなることを特徴とするスケジューリング方法。

【請求項１２】　入力ｉがｋ番目のタイムスロットに対するスケジュールを完了しなかったならば、出力を選択し、更新された出力集合を次の入力へ送出し、前記入力ｉが前記ｋ番目のタイムスロットに対するスケジュールを完了するならば、前記出力集合を前記次の入力へ送出しない、ことを特徴とする請求項１１記載の方法。

【請求項１３】　出力の更新集合を前の入力から受信しない入力は、新たなスケジュールを開始することを特徴とする請求項１２記載の方法。

【請求項１４】　入力が奇数個の場合のパイプライン・ラウンドロビン・グリーディ・スケジューリング方法において、
ｅ）ｋ（ｉ，ｈ）>0は入力ｉがｈ番目のタイムスロットにおいて確保する出力のタイムスロットを示し、i_h＝(定数－Ｎ－ｈ)modNはｈ番目のタイムスロットで新しくスケジューリングをはじめる入力を示し、ｋ（ｉ，ｈ）＝０はｈ番目のタイムスロットでの入力iの動作が抑制されることを意味するものとして、k(0,1) = k(1, 1) = ... = k(N−1,1)=0、定数=N+1に初期化し、
ｆ）O_h+N= {0,1,…,Ｎ−1}、0≦i≦Ｎ−1及びi≠i_hとして、ｋ（i_h，ｈ）＝ ｈ+N 及びｋ（ｉ，ｈ）＝ｋ（（ｉ−1）modN，ｈ−１）と設定し、
ｇ）0≦i≦Ｎ−1である入力iでパケットを送信すべき出力ｊを集合$O_k(i,h)$からラウンドロビン方式で選択し（ただしk(i,h)≠0である）、ｊを$O_k(i,h)$から除外し、
ｈ）0≦i≦Ｎ−1である入力iで選択された出力ｊのアドレスをコネクションメモリのメモリ位置k(i,h)modNに記憶し、対応する受信入出力キューからライン先頭のHOLパケットが別個の送信入出力キューへ移動し、
ｉ）0≦i≦Ｎ−1でi≠(i_h−2)modNである入力iで、次の入力(i + 1) mod Nへ集合$O_k(i,h)$を転送し、
ｊ）0≦i≦Ｎ−1である前記入力iと、入力iのメモリロケーション（hmod(N+1)）から読み込まれたアドレスの出力との間にクロスバーコネクションを確立し、
ｋ）0≦i≦Ｎ−1である各入力iについて、スケジューリングされた送信入出力キューの先頭に保持されたパケットをスイッチコアを通して送信する、
ステップからなるプロセスを用いてｈ番目のタイムスロットを完了させる方法。

【請求項１５】　入力が偶数個の場合のパイプライン・ラウンドロビン・グリーディ・スケジューリング方法において、
ｍ）ｋ（ｉ，ｈ）>0は入力ｉがｈ番目のタイムスロ

ットにおいて確保する出力のタイムスロットを示し、ｉｈ＝(定数－Ｎ－ｈ)modNはｈ番目のタイムスロットで新しくスケジューリングをはじめる入力を示し、ｋ（ｉ，ｈ）＝０はｈ番目のタイムスロットでの入力iの動作が抑制されることを意味するものとして、k(0,1) = k(1,1) = ... = k(N－1,1)=0、定数=N+1に初期化し、

ｎ）　O_{h+N+1}＝{0,1,…,N－1}、0≦ｉ≦Ｎ－1であるｉで集合｛ｉｈ，（ｉｈ＋１）ｍｏｄＮ｝の要素でないとして、ｋ（ｉｈ，ｈ）＝ｈ＋Ｎ＋１、ｋ（ｍｏｄ（ｉｈ＋１）ｍｏｄＮ，ｈ）＝ｋ（ｉｈ，ｈ－２）、及びｋ（ｉ，ｈ）＝ｋ（（ｉ－１）ｍｏｄＮ，ｈ－１）に設定し、

ｏ）　0≦ｉ≦Ｎ－1である入力iにおいてパケットを送信すべき出力ｊを集合$O_{k(i,h)}$からラウンドロビン方式で選択し（ただし、k(i,h)≠0である）、ｊを$O_{k(i,h)}$から除外し、

ｐ）　0≦ｉ≦Ｎ－1である入力iにおいて、選択された出力ｊのアドレスをコネクションメモリのメモリ位置k(i,h)mod（Ｎ＋１）に記憶し、対応する受信入出力キューからライン先頭のHOLパケットが別個の送信入出力キューへ移動させ、

ｑ）　0≦ｉ≦Ｎ－1でｉ≠(ｉｈ－2)modNである入力iにおいて、（ｉｈ－2)modNの入力が集合$O_{k((ih-2)modN,h)}$を１タイムスロット分だけ遅延させた後、次の入力(i + 1) mod Nへ 集合$O_{k(i,h)}$を転送し、

ｒ）　0≦ｉ≦Ｎ－1である入力iと、入力iのメモリ位置(ｈmod(N+1))から読み込まれたアドレスの出力との間にクロスバーコネクションを確立し、

ｓ）　0≦ｉ≦Ｎ－1である各入力iについて、スケジュールされた送信入出力キューの先頭のパケットをスイッチコアを通して送信する、

ステップからなるプロセスを用いてｈ番目のタイムスロットを完了させる方法。

【請求項１６】　マルチキャスト・スケジューリングがラウンドロビン・グリーディ・スケジューリング・アルゴリズムに組み込まれており、マルチキャストパケットは到着順に処理される方式で格納され、ユニキャストキューよりも優先的に処理され、ｈ番目のタイムスロットにおけるステップは、更に、

ｕ）　0≦ｉ≦Ｎ－1である入力iにおいて、$j \in O_{k(i,h)} \cap BM_i$を満たす全ての出力jを選択し、HOLマルチキャストパケットをk番目のタイムスロットで選択された出力へ送信し、

ｖ）　$O_{k(i,h)} \cap BM_i$が空集合であればユニキャストキューを処理し、それ以外の場合は、選択された出力を$O_{k(i,h)}$及びBM_iから除外し、

ｗ）　BM_iが空であれば、HOLマルチキャストパケットをマルチキャストキューから削除する、

ステップからなることを特徴とする請求項１４記載の方法。

【請求項１７】　マルチキャストスケジューリングがラウンドロビン・グリーディ・スケジューリング・アルゴリズムに組み込まれており、マルチキャストパケットは到着順に処理される方式で格納され、ユニキャストキューよりも優先的に処理され、ｈ番目のタイムスロットにおけるステップは、更に、

ｘ）　0≦ｉ≦Ｎ－1である入力iにおいて、$j \in O_{k(i,h)} \cap BM_i$を満たす全ての出力jを選択し、HOLマルチキャストパケットをk番目のタイムスロットで選択された出力へ送信し、

ｙ）　$O_{k(i,h)} \cap BM_i$が空集合であればユニキャストキューを処理し、それ以外の場合は、選択された出力を$O_{k(i,h)}$及びBM_iから除外し、

ｚ）　BM_iが空であれば、HOLマルチキャストパケットをマルチキャストキューから削除する、

ステップからなることを特徴とする請求項１５記載の方法。

【請求項１８】　ステージｉが入力Ｉに関連し、前記ステージｉは未来のタイムスロットでの出力への送信をスケジューリングし、前記未来のタイムスロットは全てのステージを通して順次ずれて行くＮｘＮスイッチをスケジューリングするためのＮステージ・パイプラインシステムにおいて、

入力に対応する全てのパイプラインステージは、２つの入力が同時に同一の未来のタイムスロットを選択しないように、同時にスケジューリングを実行し、出力スロットはラウンドロビン方式に基づいて選択され、出力があるステージにより選択されるとき、当該出力は、パイプラインステージが入力によってあるタイムスロットで既に選択された出力を選択しないように、出力のフリープールから除去される、ことを特徴とするＮステージ・パイプラインシステム。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】本発明は電子及び光学メディア用のアプリケーションで使用される超高速（テラビット級）交換システムに係り、特に、超高速交換システムにおける入出力間のスケジューリング及びそのスケジューリングを実現するための交換システムに関する。

【０００２】

【従来の技術】広帯域化の要求に伴い、テラビットスイッチングの必要性が益々高まってきており、このような高速スイッチングに関する提案が多くなされている。ベシャイ及びミインタ（M. Beshai and E. Miinter）による "Multi-tera-bit/s switchbased on burst transfer and independent shared buffers," ICC' 95, pp.1724-1730、マッケオウン等（N. McKeown et al）による "The tiny tera: a packet switch core," IEEE Micro, vol.171, Jan.-Feb. 1997, pp.26-33、及びゾン、シマズ、ツクダ及びユキマツ（W. D. Zhong, Y. Shimazu,

M.Tsukuda, and K. Yukimatsu) による “A modular T bit/s TDM-WDM photonic ATM switch using optical buffers,” IEICE Transactions on Communications, vol.E77-B,no.2, Feb. 1994, pp.190-196 を参照されたい。

【０００３】電子的に制御される光スイッチングコア（見方を変えれば論理的クロスバースイッチ）は高性能スイッチの候補として魅力的である。10Gbpsのライン速度で、64バイトのセル／パケットが40ns以内に処理される必要があるからである。

【０００４】当業者にとって重要な問題の１つは、このような光スイッチングコアを効率的に使用するスケジューリングの高速決定法の確立である。この場合、スイッチ（交換機）の設計において、入力バッファリング、出力バッファリングあるいはそれら両方を用いることが可能である。出力バッファリングを用いるスイッチでは、出力バッファのアクセス速度がスイッチ全体のスループットを上回っている必要がある。

【０００５】このような出力バッファに必要とされる速度を抑えるために、ノックアウト法が採用される。この方法では、有限個のセルのみが出力バッファに受け入れられ、残りは廃棄される。光ノックアウトスイッチは、ゾン、シマズ、ツクダ及びユキマツ（Zhong, Y. Shimazu, M.Tsukuda, and K. Yukimatsu）による “A modular Tbit/s TDM-WDM photonic ATM switch using optical buffers,” IEICE Transactions on Communications, vol. E77-B, no.2, Feb. 1994, pp.190-196で提案されている。光ノックアウトスイッチでは、各出力がいくつかの光逆Banyan網と光バッファを必要とするために構成が複雑となる。

【０００６】入力バッファを用いるスイッチはもっと効率的にバッファを使用でき、メモリの帯域幅もライン速度の2倍で済む。簡単な方法では、各入力がその入力の待ち行列（キュー）における最初のパケットを送信する要求を出す。もし2以上の入力が同一の出力ポートを要求した場合、そのうちの一つがランダムに選ばれる。この方法は、カロル等（M. J. Karol, M.G. Hiuchyj, and S.P. Morgan）“Input vs. output queuing on a space-division packet switch,” IEEE Transactions on Communications, vol. COM-35, no.12, Dec. 1987, pp. 1347-1356に示されており、この入力バッファリングアルゴリズムによって、均一トラフィック状況で0.587のスループットが得られている。トラフィックが均一でない場合では、スループットは更に低下する。

【０００７】他のいくつかのスケジューリング法においては、HOL（Head Of Line：ラインの先頭）パケット以外のパケットが出力ポートを要求する。ファン、アキヤマ及びタナカ(R. Fan, M. Akiyama, and Y. Tanaka)による “An input buffer-typeATM switching using schedule comparison,” Electronics and Communications in Japan: Part I, vol.74, no.11, 1991, pp.17-25; モトヤマ、ベータ及びフロスト（S. Motoyama, D.W. Petr, and V.S. Frost)による “Input-queuedswitch based on a scheduling algorithm,” Electronics Letters, vol.31, no.14, July 1995, pp.1127-1128；オバラ(H. Obara)による “Optimum architecture for input queuing ATM switches,” Electronics Letters, vol.27, no.7, March 1991, pp.555-557を参照されたい。

【０００８】より詳しく言えば、各タイムスロットにおいて、１つの入力は複数の出力ポートを要求する。タイムスロット毎にちょうど4つの要求を出せば、１に近い効率が得られる。しかしながら、この方法では、高速時、スケジューリングのための複数の要求及びその確認応答を１タイムスロット内で処理することができない（ここで１スロットはパケット送信時間を表す）。また、トラフィックに急激な変化がある場合には、各入力がどの出力を要求するかを独立して決定することからスイッチパフォーマンスは低下するであろう。

【０００９】スイッチパフォーマンスを改善したければ、スイッチコントローラがすべての入出力バッファのキューを知っていればよい。この情報によって、スイッチコントローラは各タイムスロットでの同時送信数を増加させることができる。しかしながら、ＳＬＩＰプロトコルにおいては、複数の出力が独立して複数の入力に許可を与えるので、非効率的になっている。詳しくは、N. McKeown et al. “Scheduling cells in an input-queued switch,” Electronic Letters, vol.29, no.25, Dec. 1993, pp.2174-2175を参照されたい。

【００１０】また、アルゴリズムによって入力間のより良い協調を達成している例としては、D. Guo , Y. Yemini, Z. Zhang, “Scalable high-speed protocols for WDMoptical star networks,” IEEE INFOCOM’94.を参照されたい。ただし、このアルゴリズムはスケジューリングを決定するのに多くのタイムスロットを必要とする点で不利である。

【００１１】また、良好なパフォーマンスを示すランダム・グリーディ・スケジューリング法(RGS)が提案されている。これについては、R. Chipalkatti, Z. Xhang, and A. A. Acampora, “Protocols for optical star-coupler network using WDM: performance and complexity study,” IEEE Journal on Selected Areas in Communications, vol.11, no.4, May 1993, pp.579-589、及びD. Guo, Y. Yemini,Z. Zhang, “Scalable high-speed protocols for WDM optical star networks,” IEEE INFOCOM’94を参照されたい。

【００１２】この従来のRGS法では、入力及びそれに対応するマッチングした出力の両方がランダムに選択される。しかしながら、実際には、このようなランダム選択を実装するのは困難である。各タイムスロットにおいて、N個のパケットがN個の入力からN個の出力へ転送さ

れ得ることに注意されたい。

【００１３】

【発明が解決しようとする課題】上述したように、光スイッチングコアを効率的に使用するためのスケジューリング決定方法としては未だ満足すべきものがない。すなわち、入出力バッファリングを用いるものでは、構成の複雑化やスイッチのスループットの低下を引き起こす。また、スイッチコントローラが入出力バッファのキューを集中管理する方式では、スイッチの入力数及び出力数が増大するに伴って計算量が著しく増大し、１タイムスロット内で処理を完了することが困難となる。

【００１４】そこで、本発明の目的は、超高速交換システムにおいて入出力間スケジュールを高速に確立することができる新規なスケジューリング方法及び装置（ラウンドロビン・グリーディ・スケジューリング：以下、RRGSという。）を提供することにある。

【００１５】本発明の他の目的は、交換機の内部速度を上昇させることなく厳しいタイミング要求を満たすと共に良好なパフォーマンスを得ることができる新規なパイプラインアーキテクチャを有する交換システムを提供することにある。

【００１６】

【課題を解決するための手段】本発明によるスケジューリング方法は、Ｎ入力及びＮ出力（Ｎは２以上の整数）を有し、各入力は前記Ｎ出力にそれぞれ対応したＮ個の論理的待ち行列（論理的キュー）からなる交換システムにおいて、ａ）任意のスケジューリング過程を開始する入力に対して、当該開始入力から予め定められた数のタイムスロットだけ未来のタイムスロットを決定し、ｂ）前記開始入力からラウンドロビン方式で順次１つの入力を選択し、ｃ）選択された入力に転送すべきパケットが存在すれば、当該選択された入力対して、前記未来のタイムスロットにおける利用可能な出力の集合の中から１つの出力を選択し、ｄ）選択された出力を前記出力集合から除外し、前記選択された入力とそれに関連する前記選択された出力との組をスケジュールとして記憶する、ステップからなることを特徴とする。更に、複数のスケジューリング過程がパイプライン処理により同時に進行することを特徴とする。

【００１７】Ｎが奇数の場合には、前記未来のタイムスロットは前記開始入力からＮタイムスロットだけ未来に位置し、前記選択された出力を前記出力集合から除外した出力集合を次に選択されるべき入力へ転送する。

【００１８】Ｎが偶数の場合には、前記未来のタイムスロットは前記開始入力から（Ｎ＋１）タイムスロットだけ未来に位置し、前記選択された出力を前記出力集合から除外した出力集合を次に選択されるべき入力へ転送する動作を（Ｎ＋１）タイムスロットのうち１回だけ１タイムスロット分遅延させる。

【００１９】本発明によるタイムスロット決定方法は、Ｎ入力及びＮ出力（Ｎは２以上の整数）を有し、各入力は前記Ｎ出力にそれぞれ対応したＮ個の論理的待ち行列（論理的キュー）からなるラウンド・ロビン・グリーティ・スケジューリングプロトコルのためのクロスバースイッチにおけるタイムスロット決定方法であって、前記プロトコルの入力は全ての入出力キューの状態であり、前記プロトコルの出力はスケジュールであり、

ａ）ｉ=(定数－ｋ－1)modNに対応する入力を選択し、

ｂ）もし入力が無ければ停止し、それ以外であればラウンドロビンのやり方でi = (i + 1) mod Nにより決定される次の入力を選択し、

ｃ）集合C = {(i, j) |出力jに対応する入力ｉにおいて少なくとも１個のパケットが存在する}の要素である組（ｉ，ｊ）が存在するならば、出力jを選択し、

ｄ）ステップｃ）において前記組（ｉ，ｊ）が存在しなければ、入力集合からｉを除去してステップｂ）に戻り、

ｅ）入力集合からiを、出力集合からjをそれぞれ除去し、

ｆ）前記組（ｉ，ｊ）を前記スケジュールに加えてステップｂ）に戻る、

ステップからなることを特徴とする。

【００２０】また、本発明によるスケジューリング方法は、各タイムスロットにおいて、Ｎ個の異なるスケジュールが未来のＮタイムスロット間で同時進行するスケジューリング方法において、

ａ）ラウンドロビン方式で、スケジューリングのための入力に対して特定の未来のタイムスロットを利用可能にし、

ｂ）入力ｉによって、未来のｋ番目のタイムスロットに対する出力を選択し、

ｃ）未来のｋ番目のライムスロットに対するスケジュールを開始し、

ｄ）次の入力（ｉ＋１）ｍｏｄＮを決定し、前記ｋ番目のタイムスロットの間にパケットを自由に受信できる残りの出力を前記次の入力へ送出する、

ステップからなることを特徴とする。

【００２１】本発明による入力が奇数個の場合のパイプライン・ラウンドロビン・グリーディスケジューリング方法は、

ｅ）ｋ（ｉ，ｈ）＞0は入力ｉがｈ番目のタイムスロットにおいて確保する出力のタイムスロットを示し、ｉｈ＝(定数－Ｎ－ｈ)modNはｈ番目のタイムスロットで新しくスケジューリングをはじめる入力を示し、ｋ（ｉ，ｈ）＝0はｈ番目のタイムスロットでの入力iの動作が抑制されることを意味するものとして、k(0,1) = k(1, 1) = ... = k(N−1,1)=0、定数=N+1に初期化し、

ｆ）0_{h+N}= {0,1,…,N−1}、0≦i≦N−1及びi≠iｈとして、ｋ（i_h，ｈ）＝ h+N 及びｋ（ｉ，ｈ）＝ｋ（（ｉ−1）modN，ｈ−1）と設定し、

g） 0≦i≦N−1である入力iでパケットを送信すべき出力ｊを集合$O_{k(i,h)}$からラウンドロビン方式で選択し（ただしk(i,h)≠0である）、jを$O_{k(i,h)}$から除外し、

h） 0≦i≦N−1である入力iで選択された出力ｊのアドレスをコネクションメモリのメモリ位置k(i,h)modNに記憶し、対応する受信入出力キューからライン先頭のHOLパケットが別個の送信入出力キューへ移動し、

i） 0≦i≦N−1でi≠(i_h−2)modNである入力iで、次の入力(i + 1) mod Nへ集合$O_{k(i,h)}$を転送し、

j） 0≦i≦N−1である前記入力iと、入力iのメモリロケーション（hmod(N+1)）から読み込まれたアドレスの出力との間にクロスバーコネクションを確立し、

k） 0≦i≦N−1である各入力iについて、スケジューリングされた送信入出力キューの先頭に保持されたパケットをスイッチコアを通して送信する、

ステップからなるプロセスを用いてｈ番目のタイムスロットを完了させることを特徴とする。

【００２２】更に、本発明による入力が偶数個の場合のパイプライン・ラウンドロビン・グリーディスケジューリング方法は、

m） ｋ（ｉ，ｈ）＞0は入力ｉがｈ番目のタイムスロットにおいて確保する出力のタイムスロットを示し、ｉｈ＝(定数−N−ｈ)modNはｈ番目のタイムスロットで新しくスケジューリングをはじめる入力を示し、ｋ（ｉ，ｈ）＝０はｈ番目のタイムスロットでの入力iの動作が抑制されることを意味するものとして、k(0,1) = k(1,1) = ... = k(N−1,1)=0、定数=N+1に初期化し、

n） 0_{h+N+1}= {0,1,…,N−1} 、0≦i≦N−1であるｉで集合 {i_h，（i_h＋１）ｍｏｄＮ} の要素でないとして、ｋ（i_h，ｈ）＝ｈ＋Ｎ＋１、ｋ（ｍｏｄ（i_h＋１）ｍｏｄＮ，ｈ）＝ｋ（i_h，ｈ−２）、及びｋ（ｉ，ｈ）＝ｋ（（ｉ−１）ｍｏｄＮ，ｈ−１）に設定し、

o） 0≦ｉ≦N−1である入力iにおいてパケットを送信すべき出力ｊを集合$O_{k(i,h)}$からラウンドロビン方式で選択し（ただし、k(i,h)≠0である）、jを$O_{k(i,h)}$から除外し、

p） 0≦i≦N−1である入力iにおいて、選択された出力ｊのアドレスをコネクションメモリのメモリ位置k(i,h)mod（Ｎ＋１）に記憶し、対応する受信入出力キューからライン先頭のHOLパケットが別個の送信入出力キューへ移動させ、

q） 0≦i≦N−1でi≠(i_h−2)modNである入力iにおいて、(i_h−2)modNの入力が集合$O_{k((ih-2)modN,h)}$を１タイムスロット分だけ遅延させた後、次の入力(i + 1) mod Nへ 集合$O_{k(i,h)}$を転送し、

r） 0≦i≦N−1である入力iと、入力iのメモリ位置（hmod(N+1)）から読み込まれたアドレスの出力との間にクロスバーコネクションを確立し、

s） 0≦i≦N−1である各入力iについて、スケジュールされた送信入出力キューの先頭のパケットをスイッチコアを通して送信する、

ステップからなるプロセスを用いてｈ番目のタイムスロットを完了させることを特徴とする。

【００２３】更に、マルチキャストスケジューリングがラウンドロビン・グリーディ・スケジューリング・アルゴリズムに組み込まれており、マルチキャストパケットは到着順に処理される方式で格納され、ユニキャストキューよりも優先的に処理され、ｈ番目のタイムスロットにおけるステップは、更に、

u） 0≦i≦N−1である入力iにおいて、j∈$O_{k(i,h)}$ ∩BM_iを満たす全ての出力jを選択し、HOLマルチキャストパケットをk番目のタイムスロットで選択された出力へ送信し、

v） $O_{k(i,h)}$∩BM_iが空集合であればユニキャストキューを処理し、それ以外の場合は、選択された出力を$O_{k(i,h)}$及びBM_iから除外し、

w） BM_iが空であれば、HOLマルチキャストパケットをマルチキャストキューから削除する、

ステップからなることを特徴とする。

【００２４】本発明によるＮステージ・パイプラインシステムは、ステージｉが入力Ｉに関連し、前記ステージｉは未来のタイムスロットでの出力への送信をスケジューリングし、前記未来のタイムスロットは全てのステージを通して順次ずれて行くＮｘＮスイッチをスケジューリングするためのＮステージ・パイプラインシステムであって、入力に対応する全てのパイプラインステージは、２つの入力が同時に同一の未来のタイムスロットを選択しないように、同時にスケジューリングを実行し、出力スロットはラウンドロビン方式に基づいて選択され、出力があるステージにより選択されるとき、当該出力は、パイプラインステージが入力によってあるタイムスロットで既に選択された出力を選択しないように、出力のフリープールから除去されることを特徴とする。

【００２５】本発明によれば、Ｎ入力からラウンドロビン方式で順次１つの入力を選択し、その選択された入力に対して未来のタイムスロットにおける利用可能な出力の集合の中から１つの出力を選択し、スケジュールとして記憶する。これにより、複数の出力割当動作（スケジューリング）を並列処理することが可能となり、内部速度の向上を必要とせずに厳しいタイミング基準を満たすことができる。更に、並列処理により、RGSの良好なパフォーマンスを損なわず、100%に近い使用率を達成することができる。

【００２６】

【発明の実施の形態】図１は本発明による超高速交換システムの概念的アーキテクチャを示すブロック図である。本交換システムはN×Nクロスバースイッチ１０１からなり、Ｎ本の入力ラインから受信するパケットは全て

固定長のセルである。Ｎ個の入力ポートはそれぞれＮ個のアービタを有し、更に各入力ポートはＮ個の出力ポートにそれぞれ対応したＮ個の論理キューを有する。後述するように、Ｎ個のアービタ及び各アービタに対応するＮ個の論理キューはＲＲＧＳアルゴリズムにより制御される。以下、ＲＲＧＳスケジューリングについて詳細に説明する。

【００２７】ラウンドロビン・グリーディ・スケジューリング（RRGS）

ＲＲＧＳプロトコルの入力は、全ての入出力キューの状態である。クロスバースイッチ１０１の各入力ポートをｉ（ｉ∈｛0,1,…,N−1｝）と記すと、このような入力、即ち入出力キュー、は次のような集合Cによって表すことができる。

【００２８】Ｃ ＝ ｛（ｉ，ｊ）｜出力jに対応する入力ｉにおいて少なくとも１個のパケットが存在する｝。

【００２９】ＲＲＧＳプロトコルの出力は、Ｎ個の入力をＮ個の出力に対応づけるスケジュールである。このようなスケジュールの集合Sは次のように表すことができる。

【００３０】Ｓ ＝ ｛（ｉ，ｊ）｜入力 ｉから出力jへパケットが送られる｝。

【００３１】当業者にとっては明らかであるが、各タイムスロットにおいて、１つの入力は１個のパケットのみを送信でき、１つの出力は１個のパケットのみを受信できる。この条件下で、任意のk番目のタイムスロットのスケジュールは、次のステップ１）～４）によって決定される。

【００３２】ステップ1） 全入力の集合をI_k＝｛0,1,…,N−1｝、全出力の集合をO_k＝｛0,1,…,N−1｝とする。ｉ＝(定数−k−1)modNの計算式によって、即ち（定数−ｋ−１）をＮで除算した剰余として入力ｉを選択する。スケジュールを開始する入力をこのように選択することでスケジューリングのインプリメンテーションを簡単にできるであろう。

【００３３】ステップ2） もしI_kが空であれば、停止する。空でなければ、ラウンドロビンのやり方でi ＝（i ＋ 1）mod Nとなる次の入力iを選択する。

【００３４】ステップ3） (i,j)∈ C_kであるようにO_kから出力jをラウンドロビンのやり方で選択する。もしそのような出力が存在しなければ、I_kからiを除去し、ステップ2）に戻る。

【００３５】ステップ4） I_kから入力iを、O_kから出力jを除去し、S_kに（i,j）を加え、ステップ2）に戻る。

【００３６】上記プロトコルは、明らかに上述した従来のRGS法の改良である。上述したように、従来のRGS法では、入力及びそれに対応するマッチングした出力の両方がランダムに選択されていた。しかし、実際には、このようなランダム選択を実装するのは難しい。各タイムスロットにおいて、N個のパケットがN個の入力からN個の出力へ転送され得ることに注意されたい。

【００３７】これに対して、ＲＲＳＧでは、ある与えられたタイムスロットのスケジューリングプロセスはＮ個の位相からなる。あるタイムスロットのスケジューリングの各位相において、１つの入力は当該タイムスロットでの送信のために使用可能な残存する出力の１つを選択する。後述するように（図４参照）、１つの位相は、入力モジュール(ＩＭ)からラウンドロビン(ＲＲ)アービタへ出線リクエスト（要求）を送出し、ＲＲアービタにおいてラウンドロビン方式で出力選択を行い、そしてＲＲアービタから入力モジュールＩＭへ要求応答（出線選択応答）を送出する、というステップからなる。入力が出力を選択するラウンドロビン（RR）順序は、すべての入力に対して等しいアクセスを保証するようにタイムスロット毎に循環的にシフトする。

【００３８】奇数個入力のパイプラインRRGS

10Gbpsのような高速リンク速度においては、N位相を１つのタイムスロット（64バイトのパケットサイズを仮定すると40ns）内で終了させることは出来ない。リンク速度が高速化するにつれて、従来の技術では１タイムスロットで１位相を超える処理を完了することができない。

【００３９】この問題を解決するため、本発明ではパイプライン方式を用いる。すなわち、タイムスロット毎に、未来のN個のタイムスロットにわたって、N個の異なるスケジュールが同時進行する。ある１つのスケジュールの各位相は１個の入力だけに関連する。言い換えれば、任意のタイムスロットにおいて、他の複数の入力は、他の異なる未来のタイムスロットに対するスケジュールの位相を実行することとなる。

【００４０】定義： ある未来のタイムスロットT_kのスケジュールが完了したとは、N位相全てが完了したとき、すなわち、タイムスロットT_kにおいて送信するための出力を選択するチャンス（成功／失敗に関わらず）がすべての入力に与えられたとき、をいう。

【００４１】あるスケジュールのN個の位相を完了するのにN個のタイムスロットが必要であるが、後述するようにパイプラインアプローチを用いてＮ個のスケジュールを並列計算することにより、N個の異なるスケジュールのＮ個の位相を１タイムスロット内で完了させることができる。しかし、このことはタイムスロット毎に１つのスケジュールを完了させることと結果的に等価である。

【００４２】言い換えれば、ＲＲＧＳにおいては、ある未来のタイムスロットが全ての入力に対するラウンドロビンスケジューリングに利用可能となる。すなわち、未来のｋ番目のタイムスロットT_kのスケジュールを開始する入力ｉはラウンドロビン方式で出力を選択し、（ｉ＋１）mod Nである次の入力iに対して、当該ｋ番目のタイムスロットT_kでパケットを自由に受信可能な出力ポートを示す集合O_kを送出する。

【００４３】前の入力(i－1) mod Nからk番目のタイムスロットT_kでまだ利用可能な出力ポートの集合O_kを受け取った任意の入力iは、もし可能であれば、この中から１つの出力を選択し、もしk番目のタイムスロットT_kのスケジュールが完了しなければ、更新された出力集合O_kを次の入力i = (i + 1) mod Nへ送出する。k番目のタイムスロットT_kのスケジューリングが完了したならば、更新された集合O_kは次の入力 (i + 1) mod Nへ送出されない。これにより、現タイムスロットにおいて出力集合O_kを受け取らなかった入力(i + 1) mod Nは、次のタイムスロットで新たな未来のタイムスロットのための新しいスケジューリングを開始するであろう。

【００４４】RRGSの上記ステップ（１）は、N個のタイムスロットの期間に１度、入力が出力集合O_kを送出しないことを含意している。集合O_kを送出しない入力iはk番目のタイムスロットにおける出力を選択する最後の入力である。

【００４５】定理1 入力数Nが奇数で、(定数－k)modNの入力がk －1番目のタイムスロットにおいて集合O_kを送出しないならば、(定数－k)modNの入力はk番目のタイムスロットのスケジュールを完了させる。

【００４６】（証明）上記定理１は次のことを含意する。

① 各タイムスロットにおいて、N個すべての入力は、未来のあるタイムスロットでの送信をスケジューリングする機会を有する。

② 各タイムスロットにおいて、入力は未来の１を超えないタイムスロットでの送信をスケジューリングすることができる。

③ 各タイムスロットにおいて、ある出力は１個の入力のみから送信を受け付けるようにスケジュールされうる。

【００４７】(定数－k)modNの入力ｉをk －1番目のタイムスロットにおいて集合O_kを送出しない入力と仮定すると、それ以前のN －1個の入力の各々はk番目のタイムスロットの予約をするときに集合O_kを送出しなければならない。なお、(k －1 －j) 番目のタイムスロットにおいて、（ｉ＋ｊ）modNの入力は集合O_hを次の入力へ送出しない。また、（ｉ－ｊ）modNの入力はk番目のタイムスロットのための予約を行う。このようなスケジュールは、1≦j≦(N-1) の任意のｊについて、

$\forall(1\leqq j\leqq(N-1))\quad i-j\neq i+j \bmod N \Leftrightarrow$

$\forall(1\leqq j\leqq(N-1))\quad 2j\neq 0 \bmod N \Leftrightarrow$

Nが奇数

であるならば、実現可能である。

【００４８】従って、k番目のタイムスロットのスケジュールがk －N番目のタイムスロットでの（ｉ＋１）mod Nの入力によって開始されたということは、k －N －1番目のタイムスロットでの入力iが集合O_{k-N}を送出しなかったことを意味する。

（証明終）。

【００４９】図２は、奇数入力数の５×５クロスバースイッチを用いた場合のＲＲＧＳスケジューリングの一例を示すタイミングチャートである。図２では、５つの入力Ｉ０～Ｉ４と、それら入力が出力ポートを選択するタイムスロットＴ１、Ｔ２…との関係が示されている。

【００５０】図２において、例えばタイムスロットT5で、入力I_1はタイムスロットT_{10}で送信を行うための出力ポートの選択（スケジューリング）を行い、入力I_3はタイムスロットT_8におけるスケジューリングを行っている。また、次のタイムスロットT_6では、入力I_1はタイムスロットT_8におけるスケジューリングを行っている。以下同様である。図中に参照番号２０１で例示された太い縦線は、その前の入力でスケジューリングが完了し、次の入力で新しいスケジューリングが開始することを示している。関連するタイムスロットにおいて出力を選択する最後の入力である場合には、当該入力は次の入力へ集合Ｏを送出しない。この条件はN = 5タイムスロット毎に発生するから、入力はモジュロNカウンタによって集合0を送出しないことを決定することができる。

【００５１】最後に、ｈ番目のタイムスロット（例えば現在のタイムスロット）におけるRRGSの動作を説明する。O_kはｋ番目のタイムスロットの利用可能な出力の集合を示す。ｋ（ｉ，ｈ）＞0は、入力ｉがｈ番目のタイムスロットにおいて確保する出力のタイムスロットを示し、i_h＝(定数－N－ｈ)modNはｈ番目のタイムスロットで新しくスケジューリングをはじめる入力を示す。また、ｋ（ｉ，ｈ）＝0は、ｈ番目のタイムスロットでの入力iの動作が抑制されることを意味する。スケジューラは適切な初期化が必要で、初期化期間はNタイムスロット続く。初期化が最初のタイムスロットT_1で始まると仮定すると、初期化はk(0,1) = k(1,1) =... = k(N－1, 1)=0、定数=N+1と設定することにより開始される。つまり、それ以後更新されるまで最初のNタイムスロットはすべての入力の動作が抑制される。また、パケットは、入力ポートにおいて、論理的に分離したキュー（入出力キュー）で待機すると仮定する。この入出力キューでは各出力ポートに対応して１つのキューが設けられるために、ＨＯＬ（Head Of Line：ライン先頭）ブロッキングが防止される。更に、受信入出力キュー及び送信入出力キューも設けられている。

【００５２】図１に示すシステムに即して説明すれば、つぎのように定式化できる。

【００５３】１）　O_{h+N}= {0,1,…,N－1}、0≦i≦N－1及びi≠i_hとして、ｋ（i_h，ｈ）＝ｈ＋Ｎ及び

ｋ（ｉ，ｈ）＝ｋ（(ｉ－1) modN，ｈ－1)。

【００５４】２）　0≦i≦N－1である入力iは、パケットを送信すべき出力ｊを出力集合$O_{k(i,h)}$からRR方式で選択する。ただし、k(i,h)≠0である。なお、0≦i≦

N−1及び$i \neq i_h$である入力ｉは先行するタイムスロットにおいて$(i-1) \bmod N$の入力から集合$O_k(i,h)$を受信している。選択された出力jは、受信した集合$O_k(i,h)$から除外される。

【００５５】図１に示すシステムにおいて、Ｎ個のアービタ（０，１，…，Ｎ－１）の各々は、タイムスロット毎に順次前のアービタから、利用可能な出力ポートの情報、即ち集合$O_k(i,h)$を受け取り、その利用可能な出力ポートの中から１つの出力ポートをＲＲ方式で選択する。選択された出力ポートは集合$O_k(i,h)$から除外される。こうして更新された集合$O_k(i,h)$は、後述するように次のアービタへ転送される。

【００５６】３）　$0 \leqq i \leqq N-1$である入力iは、選択された出力ｊのアドレスをコネクションメモリのメモリ位置$k(i,h) \bmod N$に記憶する。対応する受信入出力キューからライン先頭のHOLパケットが別個の送信入出力キューへ移される。

【００５７】４）　$0 \leqq i \leqq N-1$で$i \neq (i_h-2) \bmod N$である入力i（現アービタ）は、次の入力$(i+1) \bmod N$（後続するアービタ）へ集合$O_k(i,h)$を転送する。ここでは、出力ポートの使用／不使用状況を示すN bitの情報だけを転送すればよい。

【００５８】５）　続いて、クロスバースイッチ１０１は、$0 \leqq i \leqq N-1$である入力iと、入力iのメモリロケーション（$h \bmod (N+1)$）から読み込まれたアドレスの出力ポートとの間にクロスバーコネクションを確立する。

【００５９】６）　$0 \leqq i \leqq N-1$である各入力iについて、スケジューリングされた送信入出力キューの先頭（ＨＯＬ）パケットをスイッチコアに確立されたコネクションを通して選択された出力ラインへ送信する。

【００６０】偶数個入力のパイプラインRRGS

上述したように、入力ポート数が奇数個の場合、各入力は、当該入力が未来のタイムスロットのための出力ポートを選択する最後のものであるならば、更新された集合O_kを隣接入力へ転送せず、これにより当該タイムスロットのスケジューリングを完了する。しかしながら、この奇数個用のアルゴリズムを偶数の場合に直接適用すると、いくつかの入力は複数の未来のタイムスロットに対してスケジューリングを行い、逆に、他の入力は全くスケジューリングを実行しなくなってしまう。そのため、入力ポート数が偶数個のスイッチを制御するには、上述したパイプライン技法を修正する。

【００６１】上述した定理1の証明から、制御情報をブロックする代わりに遅延させることで、入力が偶数個の場合に対応できると推論できる。入力が偶数個の場合は、各入力はN個のタイムスロットで一回だけ集合O_hを次へ送らず、次のタイムスロットにおいて、当該入力は前のタイムスロットから遅れた集合O_hを送出する。入力iが遅延した集合O_hを転送するとき、現在の集合O_kは転送しない。従って、当該入力ｉはk番目のタイムスロットにおいて出力を選択する最後の入力となる。

【００６２】定理2 入力数Nが偶数で、（定数−k）modNの入力がk −2番目のタイムスロットにおける集合O_hを遅延させ、k −1番目のタイムスロットにおいて転送するならば、（定数−k）modNの入力はk番目のタイムスロットのスケジュールを完了させる。

【００６３】（証明）（定数−k）modNの入力ｉがk −2番目のタイムスロットにおける集合O_hを遅延させ、k −1番目のタイムスロットにおいて現集合O_kの代わりに遅延された集合O_hを転送するものと仮定する。この場合、（k −1 −j）番目のタイムスロットにおいて、（i +j−1）modNの入力は集合O_mを遅延させ、（i+j）modNの入力は遅延された集合O_nを次の入力へと転送する。（k −1 −j）番目のタイムスロットにおいて、（i−j）modNの入力はk番目のタイムスロットを予約し、集合O_kを転送する。ただし、Ｎが偶数で、$0 \leqq j \leqq N/2-1$とした場合、$i-j \neq i+j \bmod N$及び$i-j \neq i+j-1 \bmod N$である。

【００６４】（ｉ−Ｎ／２）ｍｏｄＮの入力は、ｋ番目のタイムスロットをどの入力も予約しないように、（ｋ−１−Ｎ／２）番目のタイムスロットにおいて集合O_kを格納する。（ｋ−２−Ｎ／２）番目のタイムスロットにおいて、（ｉ−Ｎ／２）ｍｏｄＮの入力は、ｋ番目のタイムスロットを予約する。$N/2+2 \leqq j \leqq N$として（ｋ−１−Ｎ／２）番目のタイムスロットにおいて、（ｉ−ｊ＋１）ｍｏｄＮの入力はｋ番目のタイムスロットを予約し、集合O_kを転送する。ただし、Ｎが偶数で、$i-j+1 \neq i+j \bmod N$及び$i-j+1 \neq i+j-1 \bmod N$である。

【００６５】従って、k番目のタイムスロットのスケジューリングは、入力ｉに対応するユーザｉによって終了されるまで中断することなくパイプラインによって進行する。このスケジュールが（k−Ｎ−１）番目のタイムスロットで（ｉ＋１）ｍｏｄＮのユーザ（入力）によって開始されたということは、当該入力が（k−Ｎ−２）番目のタイムスロットにおいて制御情報（即ち、集合Ｏ）の転送を遅延させたからである。

（証明終）。

【００６６】図３は、偶数入力数の４×４クロスバースイッチを用いた場合のＲＲＧＳスケジューリングの一例を示すタイミングチャートである。図３では、４つの入力Ｉ０～Ｉ３と、それら入力が出力ポートを選択するタイムスロットＴ１、Ｔ２…との関係が示されている。

【００６７】図中の太い縦線３０１は、その前の入力でスケジューリングが完了し次の入力から新しいスケジューリングを開始することを示し、斜線部３０２は制御情報Ｏが遅延していることを示す。

【００６８】上述したように、ｈ番目のタイムスロット（例えば現在のタイムスロット）におけるRRGSの動作を説明する。O_kはｋ番目のタイムスロットの利用可能な出

力の集合を示す。k（i，h）>0は、入力iがh番目のタイムスロットにおいて確保する出力のタイムスロットを示し、i_h=(定数−N−1−h)modNはh番目のタイムスロットで新しくスケジューリングを開始する入力を示す。また、k（i，h）＝0は、h番目のタイムスロットでの入力iの動作が抑制されることを意味する。スケジューラは適切な初期化が必要で、初期化期間はNタイムスロット続く。初期化が最初のタイムスロットT_1で始まると仮定すると、初期化はk(0,1)= k(1,1) = ... = k(N−1,1)=0、定数=N+２と設定することにより開始される。つまり、それ以後更新されない限り最初のNタイムスロットはすべての入力の動作が抑制される。

【００６９】図１に示すシステムに即して説明すれば、つぎのように定式化できる。

【００７０】１）　0_{h+N+1}= {0,1,…,N−1} 、0≦i≦N−1であるiが集合｛i_h，（i_h＋１）ｍｏｄＮ｝の要素でないとして、

ｋ（i_h，ｈ）＝ｈ＋Ｎ＋１、ｋ（ｍｏｄ（i_h＋１）ｍｏｄＮ，ｈ）＝ｋ（i_h，ｈ−２）、及びｋ（ｉ，ｈ）＝ｋ（（ｉ−１）ｍｏｄＮ，ｈ−１）。

【００７１】２）　0≦ｉ≦N−1である入力iは、パケットを送信すべき出力ｊを出力集合$O_{k(i,h)}$からRR方式で選択する。ただし、k(i,h)≠0である。なお、0≦ｉ≦N−1及びｉ≠i_hである入力ｉは先行するタイムスロットにおいて(ｉ−１) mod Nの入力から集合$O_{k(i,h)}$を受信している。選択された出力jは、受信した集合$O_{k(i,h)}$から除外される。

【００７２】３）　0≦i≦N−1である入力iは、選択された出力ｊのアドレスをコネクションメモリのメモリ位置k(i,h)mod（N＋１）に記憶する。対応する受信入出力キューからライン先頭のHOLパケットが別個の送信入出力キューへ移される。

【００７３】４）　0≦i≦N−1でi≠(i_h−2)modNである入力i（現アービタ）は、次の入力(i + 1) mod N（後続するアービタ）へ集合$O_{k(i,h)}$を転送する。(i_h−2) modNの入力は集合$O_{k((i_h-2)\bmod N,h)}$を１タイムスロット分だけ遅延させた後、転送する。

【００７４】５）　続いて、クロスバースイッチ１０１は、0≦i≦N−1である入力iと、入力iのメモリ位置（h mod(N+1))から読み込まれたアドレスの出力ポートとの間にクロスバーコネクションを確立する。

【００７５】６）　0≦i≦N−1である各入力iについて、スケジュールされた送信入出力キューの先頭（ＨＯＬ）のパケットをスイッチコアに確立されたコネクションを通して選択された出力ラインへ送信する。

【００７６】マルチキャスト・スケジューリング

本発明は更にマルチキャスト機能を有する。マルチキャスト・パケットは別個のキューに格納され到着順に処理される(FCFS:first come first served)。各キューは、そのHOLパケットのあて先を示すマルチキャストビットマップ(BM)を有する。最も簡単なバージョンでは、マルチキャスト・キューはユニキャスト・キューよりも優先されて処理される。ｈ番目のタイムスロットにおけるマルチキャストの場合、次の動作が付加される。

【００７７】１）　0≦i≦N−1である入力iは、j∈$O_{k(i,h)}$∩BM_iを満たす全ての出力jを選択する。HOLマルチキャストパケットは、k番目のタイムスロットにおいて、選択された全ての出力ポートへ送信される。

【００７８】２）　$O_{k(i,h)}$∩BM_iが空集合であればユニキャストキューが動作する。それ以外の場合は、選択された出力が$O_{k(i,h)}$及びBM_iから除外される。

【００７９】３）BM_iが空であれば、HOLマルチキャストパケットがマルチキャストキューから削除される。

【００８０】スイッチコントローラの実装

図４は、N×N光クロスバースイッチの制御を行うアービタ部の構成を示すブロック図である。Ｎ個の入力ポート（０）～（Ｎ−１）にそれぞれ対応して、入力モジュールIM_0～IM_{N-1}、パイプライン処理を実行するＲＲアービタコントローラARB_0～ARB_{N-1}、及びコネクションメモリM_0～M_{N-1}が設けられている。ＲＲアービタコントローラARB_0～ARB_{N-1}は、本発明によるパイプライン処理を行うためにリング上に接続されている。

【００８１】各入力モジュールＩＭは、到着したパケットを論理的に分離したキュー（受信入出力キュー）に格納する。上述したように、各受信入出力キューによって、当該入力ポートがある特定の出力ポートへパケットを出力するように関連付けられている。入力モジュールＩＭは、到着パケットの出線リクエストＲＱによって各受信キューに格納した到着パケットの管理を行うと共に、出線リクエストＲＱを関連付けられたＲＲアービタコントローラＡＲＢへ送出する。

【００８２】ＲＲアービタコントローラＡＲＢは、入力した出線リクエストＲＱに応答して、利用可能出線情報Ｏから未来のタイムスロットで利用可能な出力ポートの１つを選択し、選択された出力ポート（出線番号）ＧＲを関連付けられた入力モジュールへ戻す。上述したパイプラインの初期化プロセスによって、どのタイムスロットにおいても、ＲＲアービタコントローラＡＲＢが同じタイムスロットに対して重複して送信スケジューリングを行わないことが保証される。

【００８３】各入力モジュールＩＭは、更に、未来のタイムスロットにおいて出力ポートの予約（スケジューリング）に成功したパケットを別個の送信入出力キューとして格納する。また、ＲＲアービタコントローラＡＲＢは、対応するコネクションメモリＭの特定場所にスケジューリング結果を書きこむ。メモリのアドレスは、パケットがスケジューリングされたタイムスロットによって決定される。

【００８４】ＲＲアービタコントローラＡＲＢは、ある

タイムスロットで未だ予約されていない全ての出力ポートを示す制御情報を後続するＲＲアービタコントローラＡＲＢに通知する。より正確には、その制御情報によって予約済みの出力に対する出線リクエストが禁止される。もし制御情報を次のポートへ転送しないものがあれば、そのＲＲアービタコントローラに続く次のＲＲアービタコントローラは未来のタイムスロットで任意の出力を選択できることとなる。

【００８５】コネクションメモリに書き込まれたスケジュールに基づいて、パケットは入力モジュールからスイッチコア１０１を通して出力モジュール(OM)へ転送される。

【００８６】なお、本発明の変更や変形は、上記記載及び教示から当業者には明らかであろう。ここでは本発明のいくつかの実施形態だけを記載したが、本発明の技術的範囲から逸脱することなく、多くの変形例をなすことは可能である。

【００８７】性能比較

まず、RRGSと同程度の複雑さ(complexity)を有する他のプロトコルとの比較を行う。「複雑さ」は、１つのスケジュールを完了するのに必要な時間によって計られる。ここでは、ＨＯＬ (Head Of Line) 、I-TDMA (Interleaved TDMA)、SLIP (Iterative round-robin matching with slip)、RGS (Random greedy scheduling)、及びＲＲＧＳプロトコルの比較を行う。

【００８８】ＨＯＬプロトコルは、入力キューを用いたスイッチのための最も簡単なプロトコルである (M. Karol et al., “Input vs. output queuing on a space-division packet switch,” IEEE Transactions on Communications, vol. COM-35, no.12, Dec. 1987, pp. 1347-1356) 。各入力は適切な出力へHOLパケットの送信要求を送出する。要求された出力は、ラウンドロビン方式で入力の１つに対して送信許可を与える。次のタイムスロットで、送信許可を受けた入力はパケットを対応する出力へ向かって送出する。

【００８９】インターリーブTDMA(I-TDMA)においては、出力は入力に固定化された方法により割り当てられる (K. Bogineni et al., “Low-complexity multiple access protocols for wavelength-division multiplexed photonic networks,” IEEE Journal on Selected Areas in Communications, vol.11, no.4, May 1993, pp.590-604) 。時間がフレームに分割され、そのフレームの各タイムスロットにおいて送信スケジュールが前もって決められている。パケットは、行く先に従って別個のキューに格納され、それらのスケジュールされたタイムスロットで送信される。

【００９０】ＳＬＩＰプロトコルは、N. McKeown et al. “Scheduling cells in an input-queued switch,” Electronic Letters, vol.29, no.25, Dec. 1993, pp. 2174-2175に提案されている。各入力は、送出すべきパケットを有する場合、それらパケットの全ての送出先である出力に対して要求を出す。要求された出力は、ラウンドロビン方式で、要求を出している入力の１つに対して送信許可を与える。複数の許可を受けた入力は、ラウンドロビン方式で、許可された出力の１つを選択する。ラウンドロビン選択は、前回選択された候補者の次から開始される。

【００９１】ＲＧＳプロトコルは、ラウンドロビン選択をランダム選択に置き換えた点を除けば、RRGSと類似している (D. Guo et al., “Scalable high-speed protocolsfor WDM optical star networks,” IEEE INFOCOM’94) 。コントローラはランダムに一連の入力を選択し、それらを比較しなかった出力とランダムに比較する。高速になると、ＲＧＳは１タイムスロット内で作業を終了することが出来ない。しかしながら、本発明者等は、ランダム選択をラウンドロビン選択で置き換えた場合の影響を検査しその性能評価を行った。

【００９２】図５は、RRGS, RGS, HOL, SLIP及びI-TDMAを実装した場合、提供されたトラフィック負荷に対する平均パケット遅延をそれぞれ示すグラフである。理論特性値とシミュレーション結果とはよく一致している。

【００９３】図６（Ａ）及び（Ｂ）は、RRGS, RGS, SLIP及びI-TDMAを実装した場合、固定負荷０．８及び０．９におけるパケット遅延の相補分布関数をそれぞれ示すグラフである。プロットされた曲線はシミュレーション結果に基づく。ほとんどの負荷に対して、RRGSはSLIP及びI-TDMAよりも性能がかなり優れている。

【００９４】図７（Ａ）は、RRGS, RGS, SLIP及びI-TDMAを実装しトラフィック負荷が非単調である条件下で、入出力キュー（ｉ，ｊ）のグループＧ１での平均パケット遅延をそれぞれ示すグラフであり、図７（Ｂ）は、同条件下で異なる入出力キューのグループＧ２での平均パケット遅延をそれぞれ示すグラフであり、図８（Ａ）は、同条件下で異なる入出力キューのグループＧ３での平均パケット遅延をそれぞれ示すグラフであり、図８（Ｂ）は、同条件下で異なる入出力キューのグループＧ４での平均パケット遅延をそれぞれ示すグラフである。

【００９５】

【発明の効果】本発明は、入力バッファ付高速スイッチのためのパイプライン・ラウンドロビン・スケジューリング方法を提供する。本発明によるRRGSプロトコルでは、図５～図８に示したように、同様の複雑さにおいて、他のプロトコルよりもパケットの平均遅延時間が短い。また、パケット遅延分布が大きく広がっていない。トラフィック負荷が非単調である場合、他のプロトコルに比べると、負荷の少ないキューの遅延時間は長くなるが、負荷の大きいキューの遅延時間は遥かに短くなっている。

【図面の簡単な説明】

【図1】本発明による超高速交換システムの概念的アー

キテクチャを示すブロック図である。

【図2】奇数入力数の５×５クロスバースイッチを用いた場合のＲＲＧＳスケジューリングの一例を示すタイミングチャートである。

【図3】偶数入力数の４×４クロスバースイッチを用いた場合のＲＲＧＳスケジューリングの一例を示すタイミングチャートである。

【図4】N×N光クロスバースイッチの制御を行うアービタ部の構成を示すブロック図である。

【図5】RRGS，RGS，HOL，SLIP及びI-TDMAを実装した場合、提供されたトラフィック負荷に対する平均パケット遅延のシミュレーション結果と理論値とをそれぞれ示すグラフである。

【図6】（A）及び（B）は、RRGS，RGS，SLIP及びI-TDMAを実装した場合、固定負荷０．８及び０．９におけるパケット遅延の相補分布関数をそれぞれ示すグラフである。

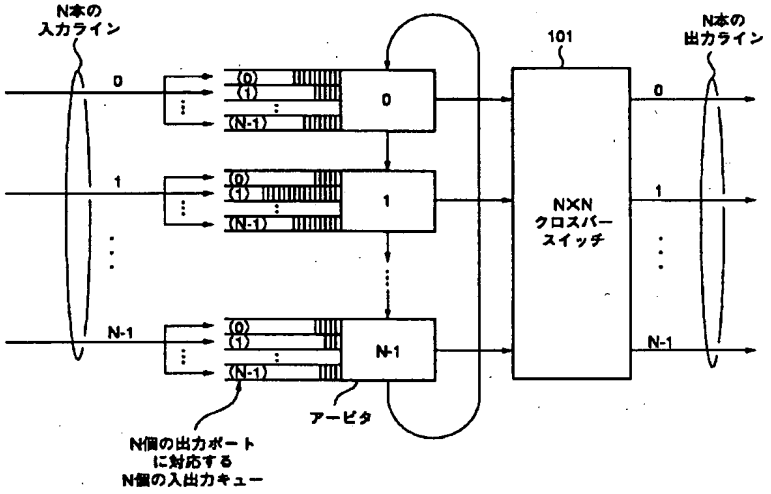
【図7】（A）は、RRGS，RGS，SLIP及びI-TDMAを実装しトラフィック負荷が非単調である条件下で、入出力キュー（ｉ，ｊ）のグループＧ１での平均パケット遅延をそれぞれ示すグラフであり、（B）は、同条件下で異なる入出力キューのグループＧ２での平均パケット遅延をそれぞれ示すグラフである。

【図8】（A）は、同条件下で異なる入出力キューのグループＧ３での平均パケット遅延をそれぞれ示すグラフであり、（B）は、同条件下で異なる入出力キューのグループＧ４での平均パケット遅延をそれぞれ示すグラフである。

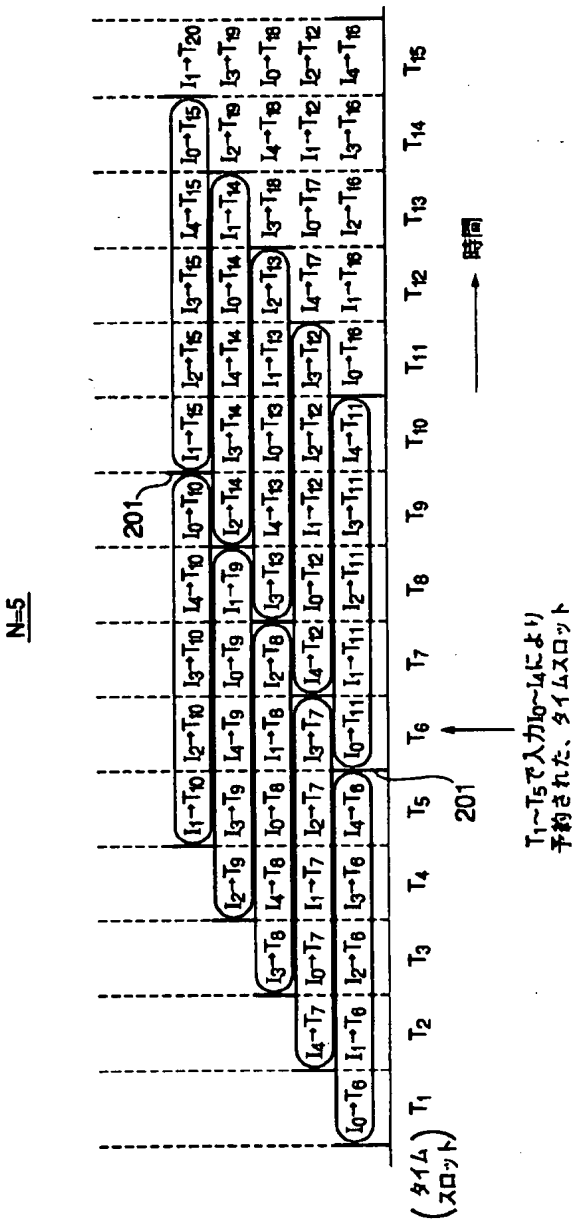
【符号の説明】

１０１　ＮｘＮクロスバースイッチ
ＩＭ　入力モジュール
ＡＲＢ　ＲＲアービタ及びパイプライン・コントローラ
Ｍ　コネクションメモリ

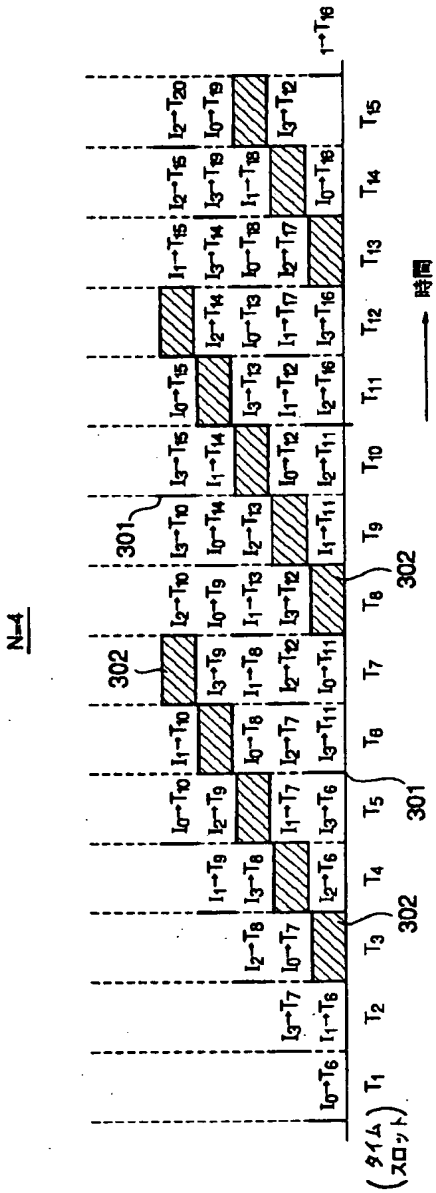
【図１】



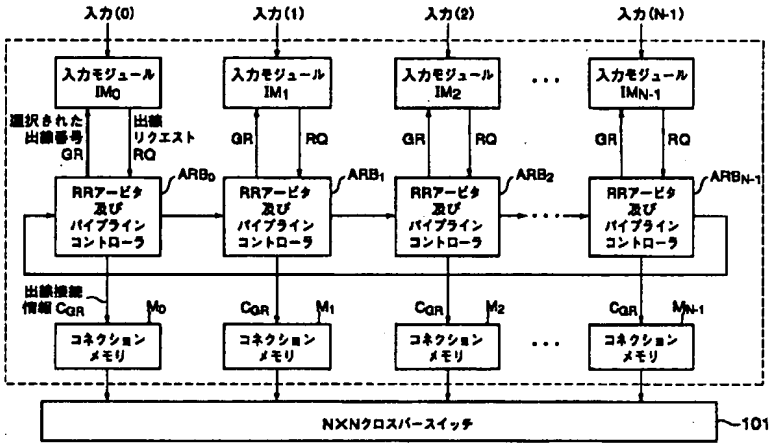
【図２】


N=5
201
時間
(タイムスロット)
T1 T2 T3 T4 T5 T6 T7 T8 T9 T10 T11 T12 T13 T14 T15
T1～T5で入力I0～I4により
予約された、タイムスロット

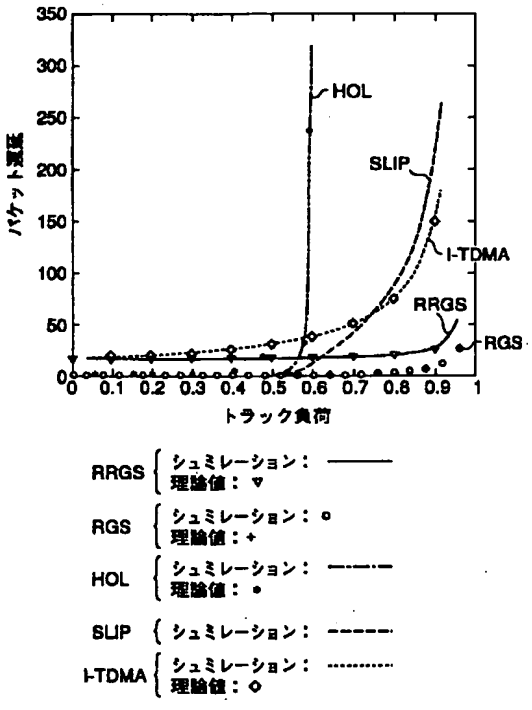
【図３】


N=4
301
302
時間
(タイムスロット)
T1 T2 T3 T4 T5 T6 T7 T8 T9 T10 T11 T12 T13 T14 T15 T16

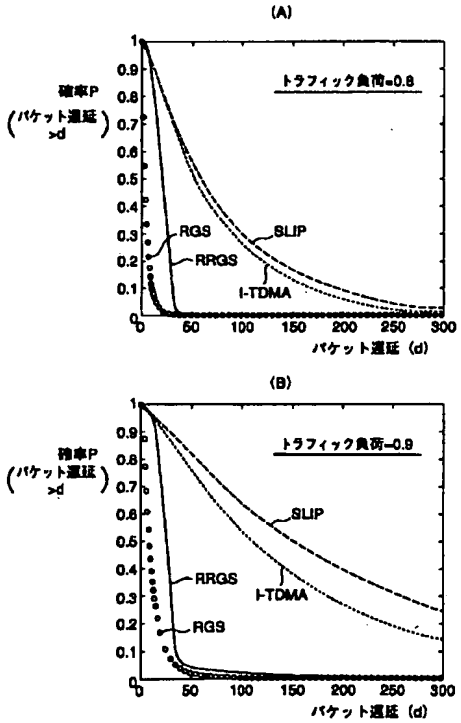
【図４】


入力(0)
入力(1)
入力(2)
入力(N-1)
入力モジュール IM0
入力モジュール IM1
入力モジュール IM2
入力モジュール IMN-1
選択された出線番号 GR
出線リクエスト RQ
GR
RQ
RRアービタ及びパイプラインコントローラ
ARB0
ARB1
ARB2
ARBN-1
出線接続情報 CGR
M0
M1
M2
MN-1
コネクションメモリ
NXNクロスバースイッチ
101

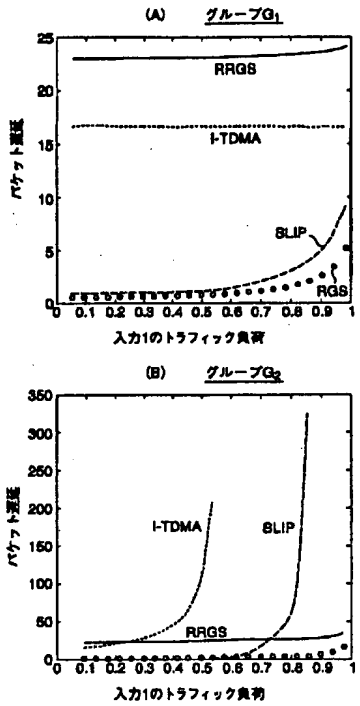
【図５】


パケット遅延
トラック負荷
HOL
SLIP
I-TDMA
RRGS
RGS
RRGS シュミレーション： 理論値：▽
RGS シュミレーション：○ 理論値：+
HOL シュミレーション： 理論値：●
SLIP シュミレーション：
I-TDMA シュミレーション： 理論値：◇

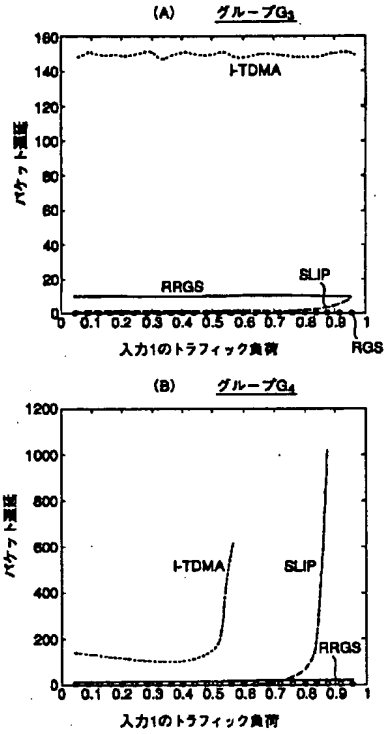
【図６】


(A)
確率P (パケット遅延 >d)
トラフィック負荷=0.8
RGS
RRGS
SLIP
I-TDMA
パケット遅延 (d)
(B)
トラフィック負荷=0.9

【図７】



【図８】



フロントページの続き

(72)発明者　ルイクシュー　ファン
アメリカ合衆国，ニュージャージー，
08540　プリンストン，インディペンデン
ス　ウエイ　４，エヌ・イー・シー・ユ
ー・エス・エー・インク内

(72)発明者　アレクサンドラ　スミルジャニク
アメリカ合衆国，ニュージャージー，
08540　プリンストン，インディペンデン
ス　ウエイ　４，エヌ・イー・シー・ユ
ー・エス・エー・インク内